REMARKS

Applicants have now had an opportunity to carefully consider the Examiner's comments set forth in the Office Action of June 4, 2007.

Reconsideration of the Application is requested in view of the amendments and comments herein. Claims 1, 11, and 20 have been amended.

The Office Action

Claims 1-20 remain in this application. Claims 1-20 are pending.

First Obviousness Rejection

The Examiner has rejected claims 1-2, 4-7, 10-12, 14-17, and 20 under 35 U.S.C. 103(a) as being unpatentable over Simske (U.S. PG Pub No. 2004/0133560) in view of Taher et al. (NPL "Evaluating Strategies for Similarity Search on the Web" ACM, May 7-11, 2002, PP 1-23) further in view of Henkin et al. (U.S. PG Pub No. 2002/0107735). This rejection should be withdrawn for at least the following reasons. Simske, Taher, and Henkin, individually or in combination, do not teach or suggest the subject invention as set forth in the subject claims.

As amended, independent claim 1 (and similarly independent claim 11) recites a a method for computing a measure of similarity between a first (or input) document and one or more disparate (or search results) documents. A first document is received and the best keywords in the text are identified by recognizing rare and uncommon keywords, including keywords that belong to one or more domain specific or subject matter specific dictionary. Documents similar to the first document are identified using a query by formulating wrappers using the list of the best keywords identified in the first document that also appear in a DS dictionary. A first list of rated keywords extracted from the first document and a list of rated keywords extracted from each of the one or more disparate documents are received. The first list of rated keywords and the list of rated keywords from each of the one or more disparate documents are compared to determine whether the first document forms part of the one or more disparate documents using a first computed percentage indicating what percentage of keyword ratings in the first list also exist in the list of at least one of the one or more disparate documents. To verify inclusion of the first document in the one or more disparate documents, a percentage is computed for each of the one or more disparate documents indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the first list also exist in the list for at least one of the one or more disparate documents when the first computed percentage indicates that the first document is included in at least one of the one or more disparate documents. The first computed percentage is used to specify the measure of similarity when the computed percentage for at least one of the one or more disparate documents is greater than the first computed percentage. The one or more disparate documents are ranked based on the percentage computed indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the first list also exist in the list for at least one of the one or more disparate documents. Simske, Taher and Henkin individually or in combination do not teach or suggest such claimed aspects of the subject invention.

In particular, Simske in view of Taher, further in view of Henkin do not teach or suggest receiving a first document and identifying the best keywords in the text by recognizing rare and uncommon keywords, including keywords that belong to one or more domain specific or subject matter specific dictionary. Simske teaches that words are weighed based on the layout of the document and overall word weight is computed primarily by counting the number of times that word occurs in the document to produce a word count. The word count is then multiplied by a “mean role weight” and a square root of the word’s lemma length to produce the total word weight.

Further, Simske in view of Taher further in view of Henkin does not teach a list of rated keywords extracted from the first document and a list of rated keywords extracted from each of the one or more disparate documents. The Examiner argues that Simske teaches this method due to Simske’s disclosure of organizing electronic documents by generating a list of weighted keywords for each document. In Simske however, keyword weight is determined by multiplying the grammatical role weight by the noun role weight and by the layout weight and then multiplying that product by the square root of the term’s lemma weight. [0028]. The present invention, keyword weight is only one way a keyword may be valued or rated. Even so, the weight determination in the present invention does not depend on the keyword’s layout in the document as it does in

Simske. The present invention assigns weight according to the rarity of a keyword, giving a higher weight to keywords that have little or no linguistic frequency.

Further, Simske does not teach or suggest computing a percentage for each of the one or more disparate documents indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the first list also exist in the list for at least one of the one or more disparate documents when the first computed percentage indicates that the first document is included in at least one of the one or more disparate documents. Simske discloses determining a mean ratio of the extended word occurrences in two documents compared to their occurrences in the larger corpus. This disclosure does not teach computing a percentage for each of the one or more disparate documents as claimed in the present invention. More specifically, Simske does not disclose using the first computed percentage to specify the measure of similarity when the computed percentage for at least one of the one or more disparate documents is greater than the first computed percentage. Simske does not disclose percentages at all, let alone use them for measuring similarity. Simske teaches computing a shared word weight for individual words to compare one document to another. There is no mention, and Simske does not contemplate using a percentage of keyword ratings in a first list to compare to a percentage of keyword ratings in a second list.

Furthermore, Simske does not teach or suggest a second percentage that is computed that indicates what percentage of keyword ratings along with a set of their neighboring keyword ratings in the first list also exist in the second list when the first computed percentage indicates that the first document is included in the second document. Instead, Simske teaches computing a "mean shared weight," which is a sum of all weight values divided by the number of documents to produce a mean value of all relevant word weights. This is not a percentage value that indicates what percentage of keyword rankings along with neighboring keyword ratings in a first list also exist in a second list when a first computed percentage indicates that a first document is included in a second document, as is recited in the subject claims. Instead, Simske teaches to an average computation that produces a mean value for the one or more word weights.

Moreover, the Examiner states that Taher teaches "ranking the one or more disparate documents indicating keyword ratings along with a set of their neighboring keyword ratings in the first also exist in the list for at least one of the one or more disparate documents;" however, the ranking method disclosed in Taher simply ranks documents according to search term matches. Taher does not rank the documents indicating keyword ratings along with a set of their neighboring keyword ratings.

With respect to independent claim 20, Applicants refer the Examiner to the above comments in reference to independent claims 1 and 11. In addition, independent claim 20 recites an article of manufacture for computing a measure of similarity between a first (or input) document and one or more disparate (or search results) document. Simske, in view of Taher, further in view of Henkin do not individually or in combination teach or suggest the claimed aspects of the subject invention.

In particular, Simske does not teach or suggest a fourth computed percentage to specify the measure of similarity except when: (i) the fourth computed percentage is greater than the second computed percentage; (ii) the first list of rated keywords is identified using OCR; (iii) the fourth computed percentage is greater than fifty percent; and (iv) less than twenty percent of the keywords in the first list of keywords are in the second list of keywords are in the second list of keywords. The Examiner states that Simske discloses that if any documents being considered are paper-based, tools such as a zoning analysis engine in combination with an optical character recognition (OCR) engine may be used to convert the paper-based document to an electric document. (Simske [0016]). This disclose does not teach or suggest a fourth computer percentage to specify the measure of similarity between the first rated keywords and the list of rated keywords for the second document. The fact that Simske identifies its keywords in the documents with OCR works to make the present invention more nonobvious since Simske did not contemplate the need for a fourth computed percentage if a third percentage indicates that the first document is a revision of a second document.

For at least the aforementioned reasons, Simske in view of Taher, further in view of Henkin individually or in combination do not teach or suggest the subject invention as recited in independent claims 1, 11, 20 (or claims 2-10 and 12-19 which respectively depend therefrom)

Second Obviousness Rejection

The Examiner has rejected claims 3 and 13 under 35 U.S.C. 103(a) as being unpatentable over Simske (U.S. PG Pub No. 2002/0107735) in view of Taher et al. (NPL "Evaluating Strategies for Similarity Search on the Web" ACM, May 7-11, 2002, PP 1-23), further in view of Henkin et al. (U.S. PG Pub No. 2002/0107735) as applied to claims 1-2, 4-7, 10-12, 14-17, and 20 above, further in view of Kubota (U.S. Patent No. 6,041,323). This rejection should be withdrawn for at least the following reasons. Claims 3 and 13 depend from independent claims 1 and 11 respectively, and the combination of references cited by the Examiner and referred to above, do not make up for the aforementioned deficiencies of Simske, Taher, and Henkin regarding the present application. Simske in view of Taher, further in view of Henkin do not teach or suggest comparing a first list of rated keywords extracted from each of the one or more disparate documents that are received, wherein the one or more disparate documents are ranked based on the percentage computed indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the first list also exist in the list for at least one of the one or more disparate documents. Thus, for at least the reasons discussed above with respect to claims 1, 11 and 20, the combination of Simske, Taher, Henkin and Kubota do not teach or suggest the subject claims. Claims 1 and 11 are now in condition for allowance; therefore, the rejection of claims 3 and 13 should be withdrawn.

Third Obviousness Rejection

The Examiner has rejected claims 9 and 19 under 35 U.S.C. 103(a) as being over Simske (U.S. PG Pub No. 2004/0133560) in view of Taher (NPL "Evaluating Strategies for Similarity Search on the Web" ACM, May 7-11, 2002, PP 1-23), further in view of Henkin et al. (U.S. PG Pub No. 2002/0107735) as applied to claims 1-2, 4-7, 10-12, 14-17, and 20 above, in view of Drissi et al. (U.S. PG Pub No. 20003/0149689). This rejection should be withdrawn for at least the following reasons. Claims 9 and 19 depend from independent claims 1 and 11 respectively. The combination of Simske, Taher and Henkin do not teach or suggest the subject invention as is described above. Claims 1 and 11 are now in condition for allowance, therefore for at least the reasons discussed above, the rejection of claims 9 and 19 should be withdrawn.

Fourth Obviousness Rejection

The examiner has rejected claims 8 and 18 under 35 U.S.C. 103(a) as being unpatentable over Simske (U.S. PG Pub No. 2004/0133560). This rejection should be withdrawn for at least the following reasons. Claims 8 and 18 depend from independent claims 1 and 11 respectively and, as noted above, Simske does not teach or suggest comparing a first list of rated keywords extracted from a first document and a list or rated keywords and comparing the received extracted from each of the one or more disparate documents that are received, wherein the one or more disparate documents are ranked based on the percentage computed indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the first list also exist in the list for at least one of the one or more disparate documents. Thus, for at least the reasons discussed above with respect to claims 1, 11 and 20, Simske does not teach or suggest the subject claims. Accordingly, the rejection of claims 8 and 18 should be withdrawn.

CONCLUSION

For the reasons detailed above, it is submitted all claims remaining in the application (Claims 1-20) are now in condition for allowance. The foregoing comments do not require unnecessary additional search or examination.

No additional fee is believed to be required for this Amendment C. However, the undersigned attorney of record hereby authorizes the charging of any necessary fees, other than the issue fee, to Xerox Deposit Account No. 24-0037.

In the event the Examiner considers personal contact advantageous to the disposition of this case, he/she is hereby authorized to call Mark Svat, at Telephone Number (216) 861-5582.

Respectfully submitted,

FAY SHARPE LLP

11/2/07
Date

Mark S. Svat, Reg. No. 34,261
Kevin M. Dunn, Reg. No. 52,842
1100 Superior Avenue, Seventh Floor
Cleveland, OH 44114-2579
216-861-5582

CERTIFICATE OF MAILING OR TRANSMISSION

I hereby certify that this correspondence (and any item referred to herein as being attached or enclosed) is (are) being

□ deposited with the United States Postal Service as First Class Mail, addressed to: Mail Stop Amendment, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on the date indicated below.

☒ transmitted to the USPTO by electronic transmission via EFS-Web on the date indicated below.

Express Mail Label No.:	Signature: Elaine M. Checovich
Date: 11-2-07	Name: Elaine M. Checovich